CDF offline operations

Rick Snider

DOE Tevatron Operations Review

March 27, 2007

Outline

- CDF computing model
- Operations
- Strategies, on-going work, improvements
- Grid computing
- Summary

CDF computing model

- From detector to physics results:
  - Raw data written to tape
  - Measure final calibrations (~4-6 weeks of data per cycle)
    - Process fraction of data stream
  - Full event reconstruction on "production farm"
  - Centralized ntuple production
    - Supports the vast majority of physics analyses
  - Monte Carlo generation/reconstruction
    - Follows data taking
  - Physicist analysis of ntuples (and production output in some cases)
    - Create smaller ntuples, specialized datasets or ntuples
  - Final analysis on user desktops / institutional machines

CDF computing model

- From detector to physics results:
  - Raw data written to tape
  - Measure final calibrations (~4-6 weeks of data per cycle)
    - Process fraction of data stream
  - Full event reconstruction on "production farm"
  - Centralized ntuple production
    - Supports the vast majority of physics analyses
  - Monte Carlo generation/reconstruction
    - Follows data taking
  - Physicist analysis of ntuples (and production output in some cases)
    - Create smaller ntuples, specialized datasets or ntuples
  - Final analysis on user desktops / institutional machines

Centrally managed computing (about 65% of total CPU usage)

CDF computing model

- From detector to physics results:
  - Raw data written to tape
  - Measure final calibrations (~4-6 weeks of data per cycle)
    - Process fraction of data stream
  - Full event reconstruction on "production farm"
  - Centralized ntuple production
    - Supports the vast majority of physics analyses
  - Monte Carlo generation/reconstruction
    - Follows data taking
  - Physicist analysis of ntuples (and production output in some cases)
    - Create smaller ntuples, specialized datasets or ntuples
  - Final analysis on user desktops / institutional machines

User-managed computing

CDF computing model

- Major hardware resources
  - Production farm (Fermilab)
    - Calibrations, full event reconstruction of all data
  - Central Analysis Farm (CAF) (Fermilab)
    - Ntuple production, user analysis, MC generation
    - 1 dedicated farm + 1 CDF-purchased farm in Fermigrid (OSG) + opportunistic use of other Fermigrid computing elements (eg, CMS)
  - Distributed CAFs (dCAF's)
    - 11 remote sites — 4 in shared pools accessed via grid mechanisms
  - Grid computing (dedicated + opportunistic OSG/LCG)
    - MC generation
  - Tape archive + disk cache (+ data handling system)
  - Databases + networks + static "project" disk

CDF computing model

- Major hardware resources
  - Production farm (Fermilab)
    - Calibrations, full event reconstruction of all data
  - Central Analysis Farm (CAF) (Fermilab)
    - Ntuple production, user analysis, MC generation
    - 1 dedicated farm + 1 CDF-purchased farm in Fermigrid (OSG) + opportunistic use of other Fermigrid computing elements (eg, CMS)
  - Distributed CAFs (dCAF's)
    - 11 remote sites — 4 in shared pools accessed via grid mechanisms
  - Grid computing (dedicated + opportunistic OSG/LCG)
    - Includes both dedicated and opportunistic resources (more on this later...)
  - Tape archive + disk cache (+ data handling system)
  - Databases + networks + static "project" disk

Production farm, CAF, dCAFs use same job submission, management software

CDF computing model

- Major hardware resources
  - Production farm (Fermilab)
    - Calibrations, full event reconstruction of all data
  - Central Analysis Farm (CAF) (Fermilab)
    - Ntuple production, user analysis, MC generation
    - 1 dedicated farm + 1 CDF-purchased farm in Fermigrid (OSG) + opportunistic use of other Fermigrid computing elements (eg, CMS)
  - Distributed CAFs (dCAF's)
    - 11 remote sites — 4 in shared pools accessed via grid mechanisms
  - Grid computing (dedicated + opportunistic OSG/LCG)
    - Includes both dedicated and opportunistic resources (more on this later…)
  - Tape archive + disk cache (+ data handling system)
  - Databases + networks + static “project” disk

Can run any production job on CAF if more CPU is needed.

CDF computing model

- Major hardware resources
  - Production farm (Fermilab)
    - Calibrations, full event reconstruction of all data
  - Central Analysis Farm (CAF) (Fermilab)
    - Ntuple production, user analysis, MC generation
    - 1 dedicated farm + 1 CDF-purchased farm in Fermigrid (OSG) + opportunistic use of other Fermigrid computing elements (eg, CMS)
  - Distributed CAFs (dCAF's)
    - 11 remote sites — 4 in shared pools accessed via grid mechanisms
  - Grid computing (dedicated + opportunistic OSG/LCG)
    - Includes both dedicated and opportunistic resources (more on this later…)
  - Tape archive + disk cache (+ data handling system)
  - Databases + networks + static "project" disk

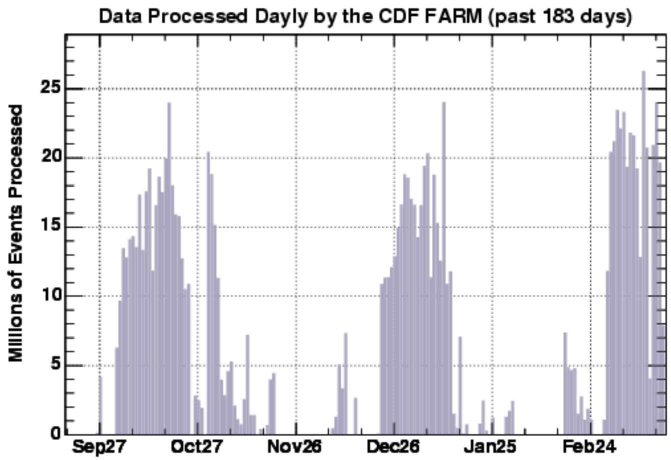
Common DH solution for CDF, D0 provided by CD

Status of operations

Production cycle operations

Operating under single-pass, pipelined scheme

- Detector requires re-calibrations every 4-6 weeks
  - Calibration data through prod. farm within 3 days of data taking
  - Final calibrations produced within ~2-3 weeks of end of period
- Process run period on production farm, CAF when calibrations are ready
  - Recently processing 20-25 M events/day
  - Takes about 2-3 weeks to complete for one run period



Data ready for physics analysis in at most 6-8 weeks after data taking

Production cycle operations

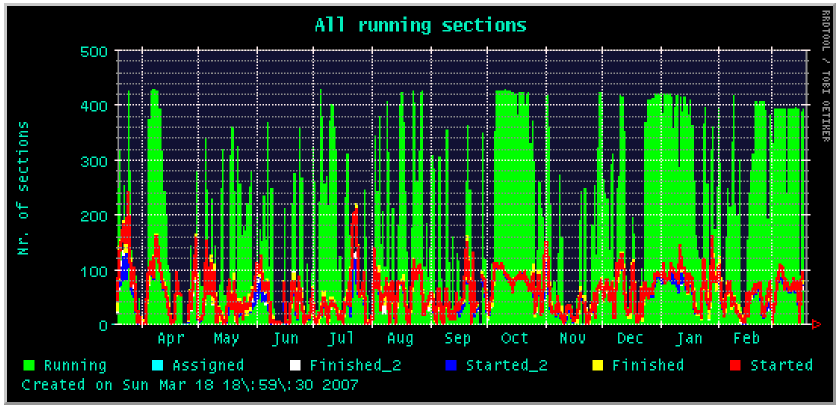
Data accumulated	Period Integ Lum	Total Integ Lum	Available for Physics
Dec 04 - Mar 19, 2005	130	680	July 2005
Mar 19 - May 20, 2005	130	810	Aug 2005
May 20 - Jul 20, 2005	100	910	Sept 2005
Jul 20 - Sep 04, 2005	95	1005	Oct 2005
Sep 5 – Nov 9, 2005	135	1140	Jan 2006
Nov 10 – Jan 14, 2006	110	1250	Mar 2006
Jan.15 - Feb.22, 2006	50	1300	Apr 2006
Jun. 9 – Sep. 01, 2006	210	1510	Oct 3, 2006
Sep. 1 – Nov. 22, 2006	180	1690	Jan 27, 2007
Nov. 24 – Jan. 30, 2007	280	1970	Ready within days

Have sustained ~6 week turn-around over almost 2 years

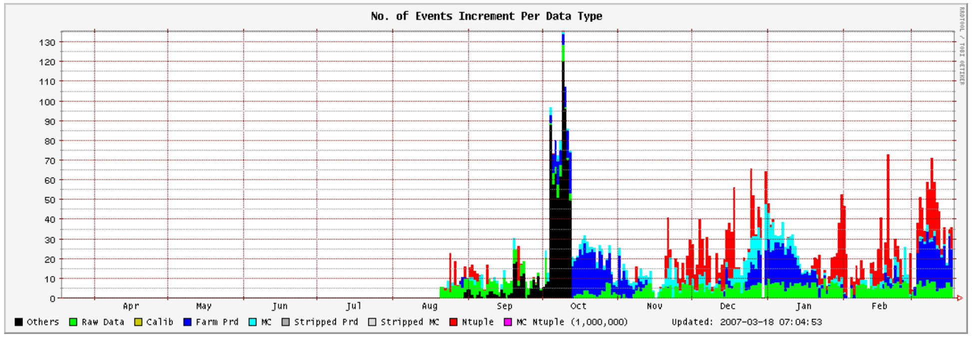
Production operations

Running jobs on production farm during past year.

Most activity follows production cycle.



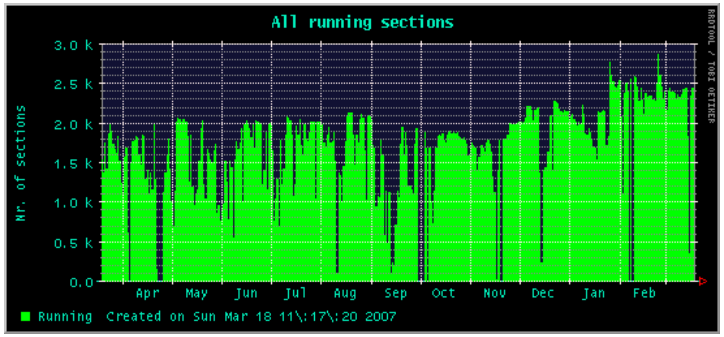
N events/day (M)

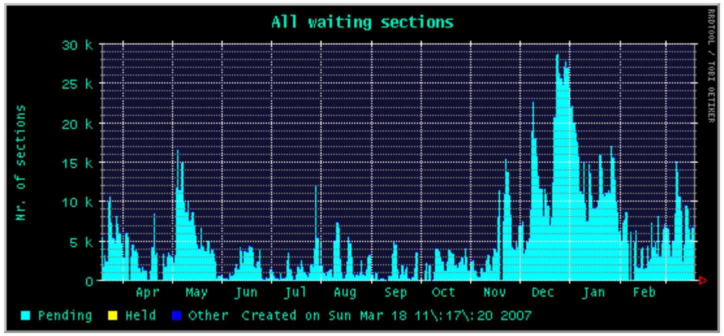


Data written by various production activities

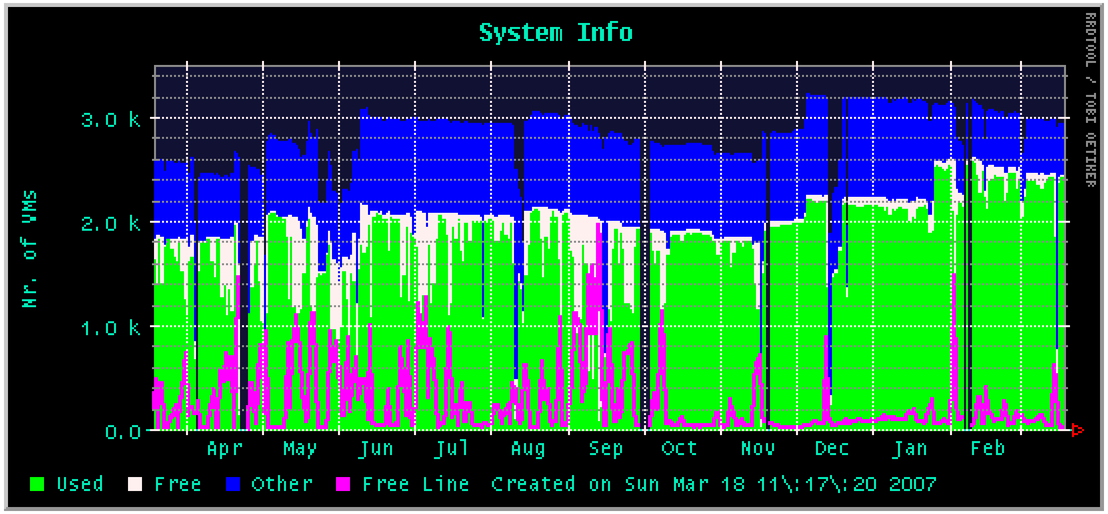
(+ 5-7 M raw data events/day)

CAF operations





Almost always jobs waiting to run.
CPU utilization is high, though not 100%.



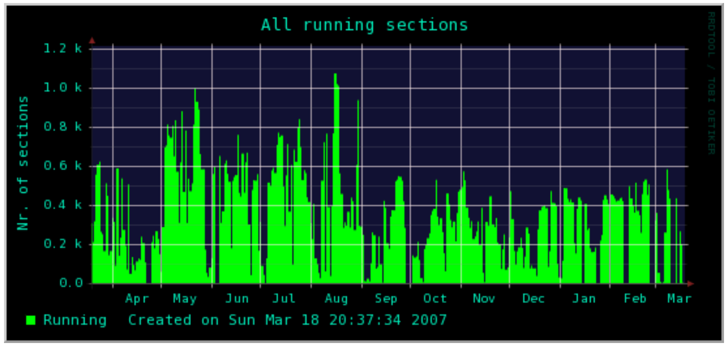
CPU usage:
- ~30% MC
- ~10% ntuplizing
- ~10% ntuple analysis
- ~50% other analysis

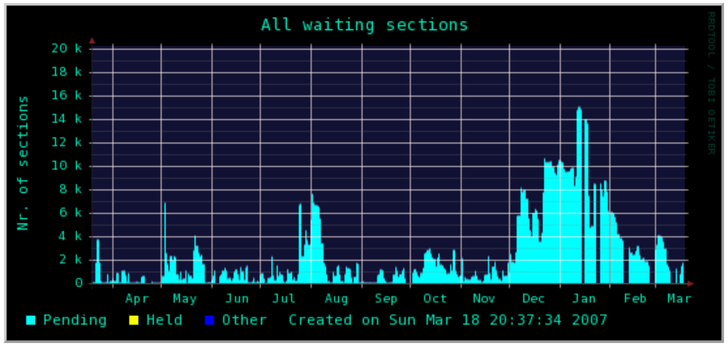
Fermigrid operations

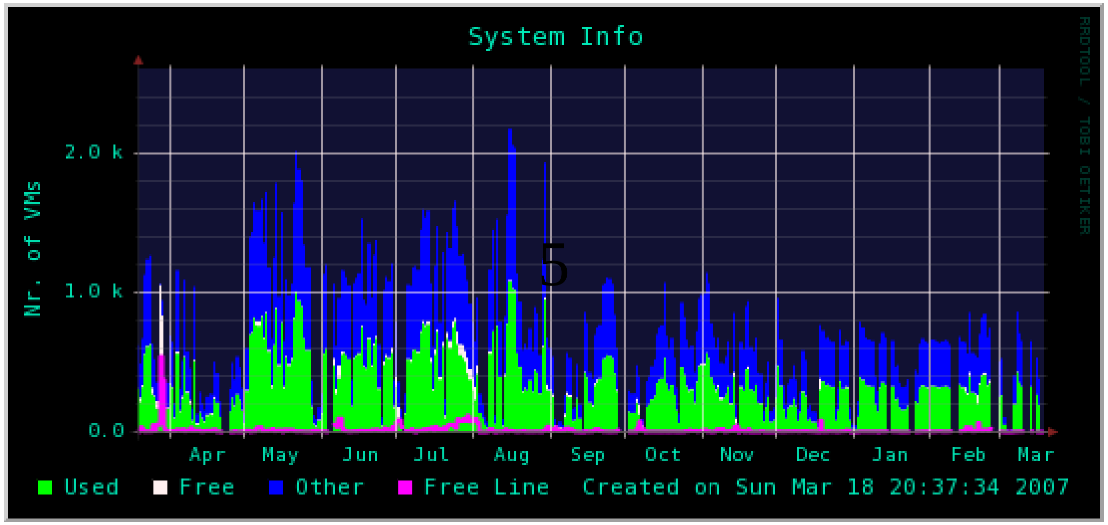
Also almost always busy.

Variation in load: "Fermigrid" shares resources with CDF owned OSG site.

CDF policy: data analysis submitted directly to "Fermigrid", MC to OSG



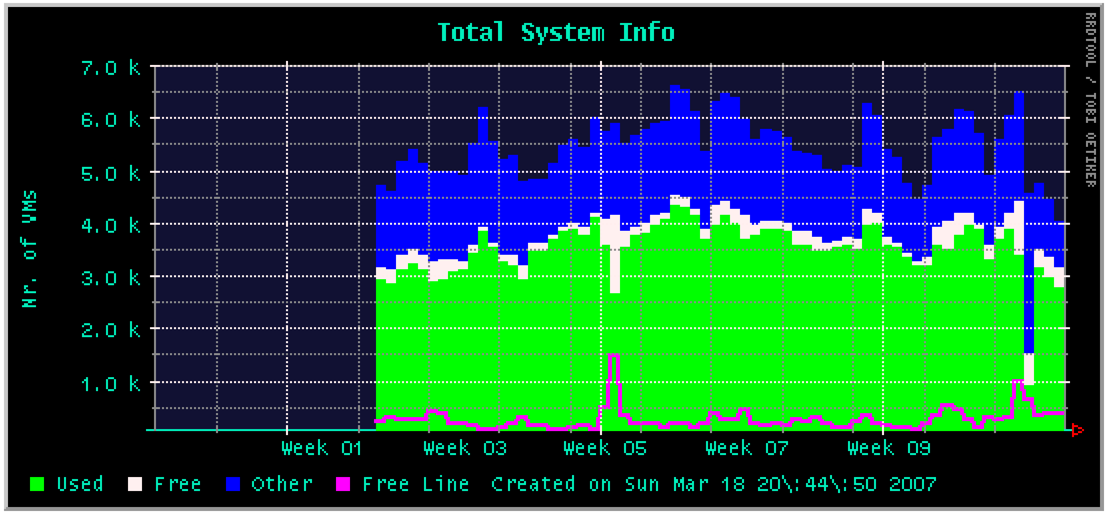




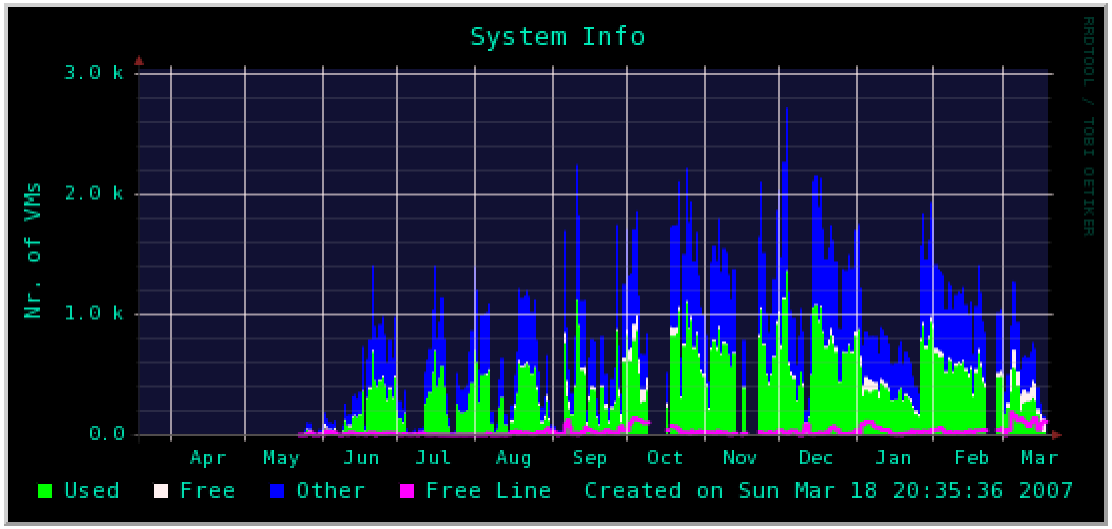
CPU usage:
- ~40% MC
- ~2% ntuplizing
- ~10% ntuple analysis
- ~50% other analysis

Remote computing operations

Sum over 11 dCAF sites



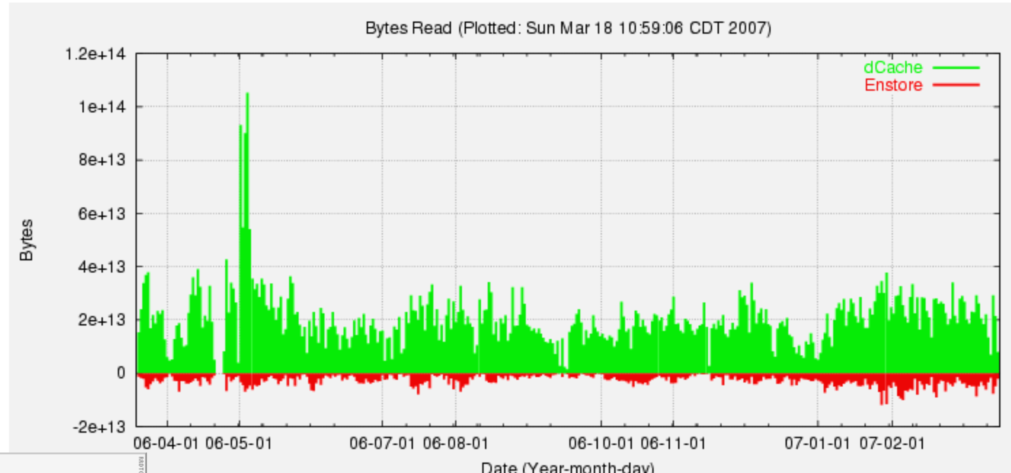
Sum over OSG sites used by CDF



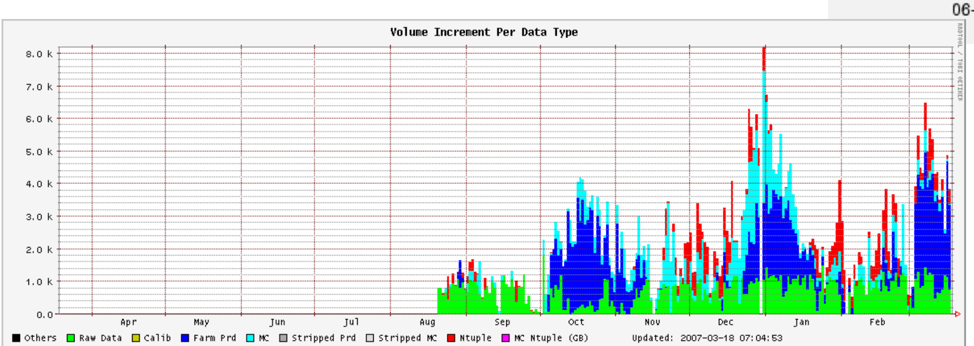
Remote CPU usage:
Typically 60-80% MC

Data handling operations

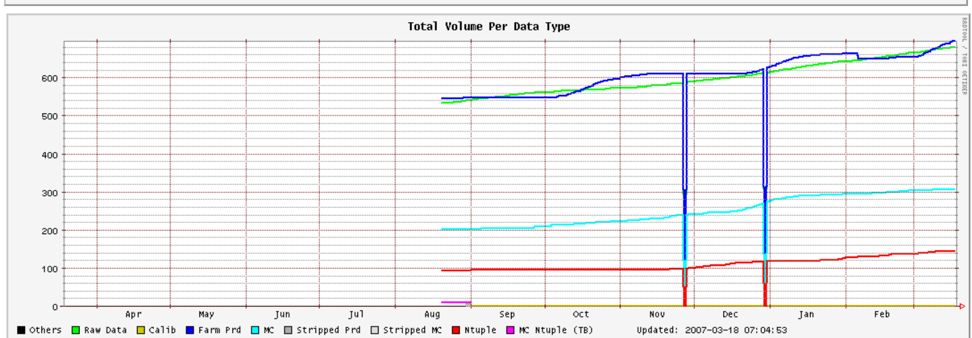
Smooth data movement and storage underlies almost all other computing activities



Data/day (GB)



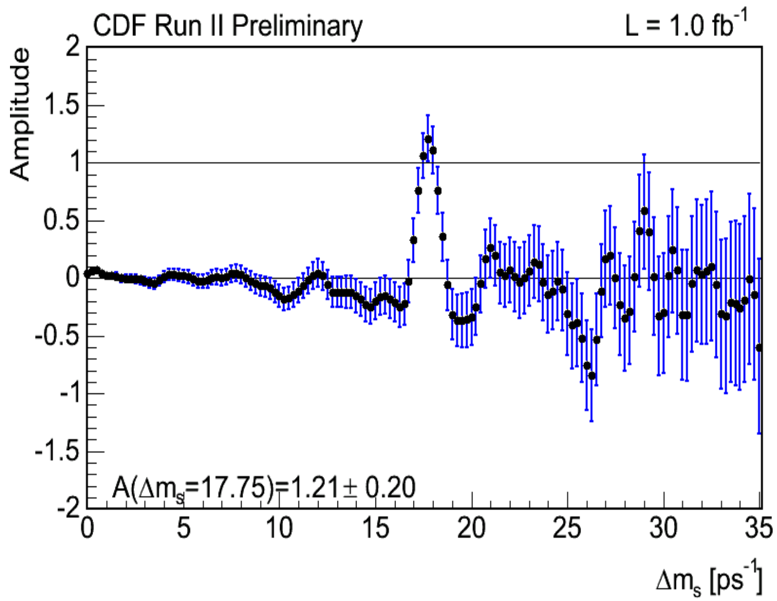
Total data vol (TB)

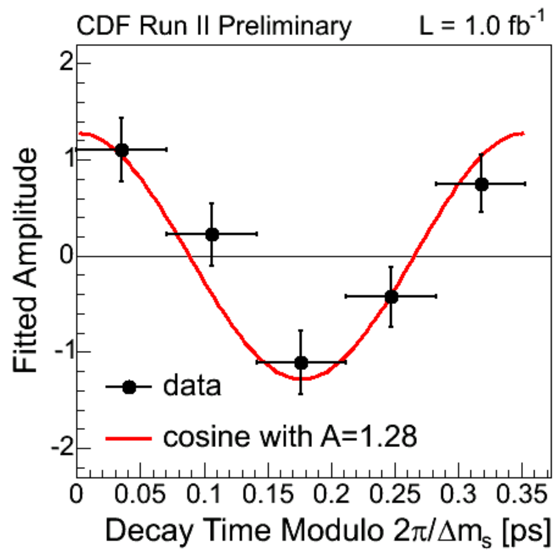


Daily and integrated data storage by production activity.

Providing data for physics

- Observation of B_s mixing
  - 8×10^{-8} probability (> 5σ) random fluctuation would look like a signal
  - $\Delta m_s = 17.77 \pm 0.01 \pm 0.07\ ps^{-1}$
  - $| V_{td} / V_{ts} | = 0.2060 \pm 0.0007$ (expt.) ± 0.0081 (theo.)





A complex analysis that uses several of the largest datasets

Collected 1 fb^{-1} by Nov. 2005

Preliminary result April 10, 2006

Final result Sept. 1, 2006

Providing data for physics

- ICHEP, July 2006
  - CDF presented ~30 new 1 fb^{-1} results
    - QCD, B, EWK, Top and Exotics

- Bs mixing: semi-lept
- Bs mixing: hadr
- Bc->J/Psi pi
- Lb lifetime
- Chic x-sections
- Orb. exc. B_s
- etab->J/PsiJ/Psi

- incl. Jet kt
- incl. Jet cone
- Kt distributions of particles in jets

- mtop l+jets
- mtop dilepton
- mtop hadronic
- top x-sec hadronic
- W helicity

- Z Pt
- ZZ

- Higgs combination
- WH->lvbb
- ZH->llbb
- ZH->vvbb

- diphoton
- HT emu
- Triphoton
- diphoton+met
- l+gamma+X
- ttbar+gamma
- monojet

- Offline is effective at meeting physics goals of experiment

Computer security

- All CDF systems comply with Fermilab Policy on Computing and with guidelines of the Fermilab Computer Security Team
- One reportable computer security incident involving CDF nodes since Jan. 1, 2006
  - Did not involve CDF online systems or Major Application
  - No impact on CDF online systems or data taking
  - No impact on data, physics results, on-going operations (details)

Evolution and improvements

Context for the computing model

- Drivers for evolution of the computing model
  - Anticipated increases in raw data logging rate
    - Capability has tripled over past year
    - Drives increasing complexity of computing with time
  - Limited budgets
    - Computing is resource constrained
  - Evolving grid infrastructure, resources
  - Anticipated decline in available effort starting around FY2007

Computing is not static

Evolution of the computing plan

- Strategy
  - Optimize use of computing resources
    - Highly centralized, incremental processing model
  - Streamline and automate operational procedures
    - Reduce effort required to produce physics results
  - Exploit emerging grid-based technologies, resources
    - Allows continued expansion of off-site computing resources
      - Currently about 25% of total CPU usage
      - Expected to grow to 40%+ in FY2008 and FY2009
    - Leverage effort devoted to (OSG/LCG) grid development
    - Distributes support load
    - Provides access to opportunistic resources

Optimizing resource utilization

- Increasingly incremental computing model in past few years
  - Production cycle has fixed CPU cost per event logged
    - CPU demand is proportional to logging rate — including MC
  - Benefits
    - Maximizes efficiency of CPU resources
    - Centrally manage resource utilization
      - Eliminate duplication
      - Promote widespread use of automation
        - Use common production infrastructure
        - Reduces effort and error rate
      - Common solutions simplify identifying, adopting efficiency improvements
        - Continuing to make improvements
  - Have centralized as much of processing as possible

Optimizing resource utilization

- Automated site selection (new development within last year)
  - Single access points for grids in North America, Europe, Pacific Rim
    - Simplifies job submission for users
    - Shields users from changes in available sites, downtimes, etc.
  - Improves load balancing across available sites
    - Allowing users-based balancing not as effective
  - Eventually all resources to be accessed via small number of portals
- Consolidation of resources at Fermilab
  - All CDF-purchased CPU at Fermilab to be within OSG pools
    - No dedicated production farm
      - Make better use of CPU within farm
    - Uniform interfaces for users
    - Single infrastructure for job submission, management

Optimizing resource utilization

- Generation 7 reconstruction
  - To be released this summer
    - Expect (hope) this to be the last production release for the experiment
  - Main focus is improvements in tracking performance
    - High luminosity + forward tracking
  - Moves a CPU-intensive part of ntuple creation into first phase of reconstruction
    - Currently some tracking improvements run during ntupling
    - Gen 7 eliminates this duplication at the ntuple creation phase
  - Will require re-processing of entire dataset

Streamline, automate operational procedures

- Automated calibration procedure (new within the past year)
  - Previously a labor-intensive effort of many detector experts
  - Developed automated framework
    - Operated by single person on production farm
    - Expert input required only at decision points
  - Significantly reduced effort required to produce calibrations
    - Expect further reductions in turn-around time with future improvements
- Use production reconstruction framework for ntuples
  - Create ntuples used by the majority of the collaboration
  - Operated by a few volunteers from physics groups
    - In principle a non-expert procedure

Streamline, automate operational procedures

- MC production
  - Dataset creation coordinated across physics groups
    - Eliminate needless replication of datasets
  - Uses common infrastructure, configurations, procedures
    - Highly automated
    - Available and used by individual users as well
    - All users benefit from centrally validated configurations, executables
  - Standard MC datasets generated to replicate each real data run
    - Demand for MC scales with logging rate or incremental luminosity

Always seeking other improvements

- Trying to make systems more robust, simple to use to reduce operational load

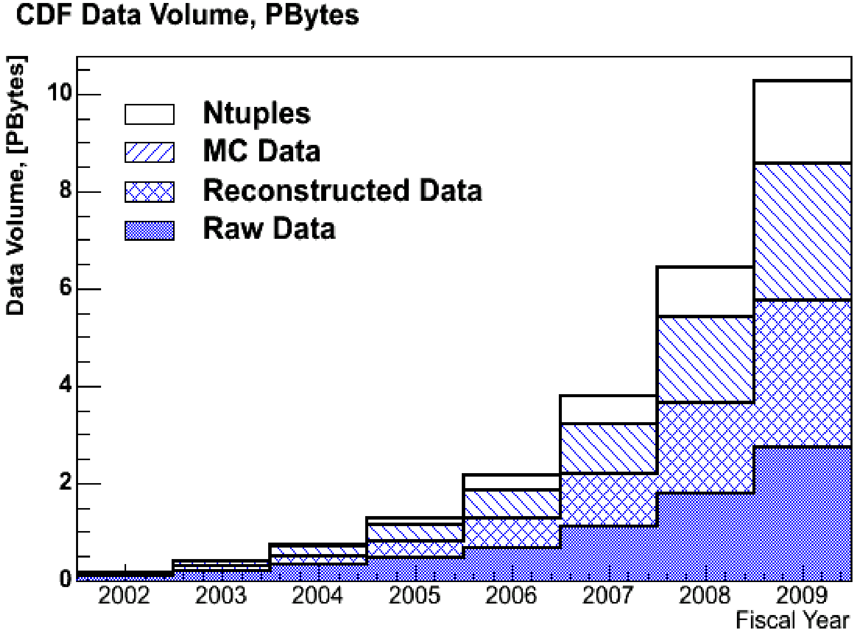
Grid computing

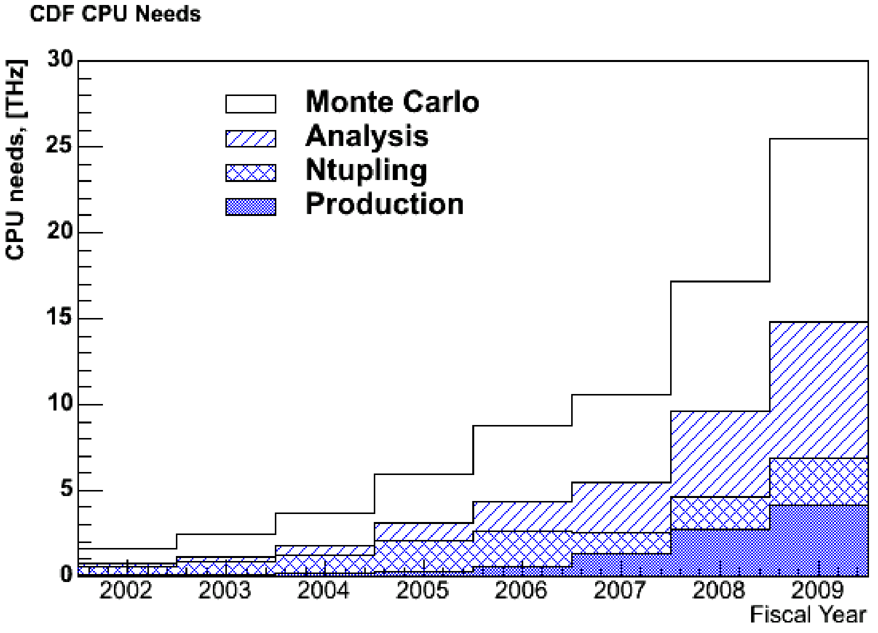
Computing model input parameters

Fiscal Year	2007	2008	2009
Integrated luminosity (fb^{-1})	3.9	5.9	8.1
Total number of events (10^9)	5.7	9.9	14
Raw data logging rate (MB/s)	30	45	60

Follows possible improvements in trigger algorithms

Worst case in terms of computing





Computing inventory

		Actual		Requirements		
	Fiscal Year	2006	2007	2007	2008	2009
CPU (THz)	Fermilab	4.8	7.6	7.6	7.8	7.9
	On-site contributions	1.7	1.7	1.7	1.7	1.7
	Remote (dedicated)	2.3	2.3	2.3	2.3	2.3
	Opportunistic		1.0	1.0	5.0	13
	Total	8.8	13	13	17	25
Disk (PB)	Fermilab	0.6	0.6	0.6	0.9	1.6
	On-site contributions	0.2*	0.1	0.1	0.1	0.1
	Remote	0.1?	0.1?			
	Total	0.9?	0.9?	0.7	1.0	1.7
Volume on tape, Oct. 1, 2006 (PB)		1.5	1.6	3.8	6.4	10.3

* 50% retires at end of FY2006

Computing inventory

Will need opportunistic resources to meet the expected demand for CPU

		Actual		Requirements		
	Fiscal Year	2006	2007	2007	2008	2009
CPU (THz)	Fermilab	4.8	7.6	7.6	7.8	7.9
	On-site contributions	1.7	1.7	1.7	1.7	1.7
	Remote (dedicated)	2.3	2.3	2.3	2.3	2.3
	Opportunistic		1.0	1.0	5.0	13
	Total	8.8	13	13	17	25
Disk (PB)	Fermilab	0.6	0.6	0.6	0.9	1.6
	On-site contributions	0.2*	0.1	0.1	0.1	0.1
	Remote	0.1?	0.1?			
	Total	0.9?	0.9?	0.7	1.0	1.7
Volume on tape, Oct. 1, 2006 (PB)		1.5	1.6	3.8	6.4	10.3

* 50% retires at end of FY2006

Computing inventory

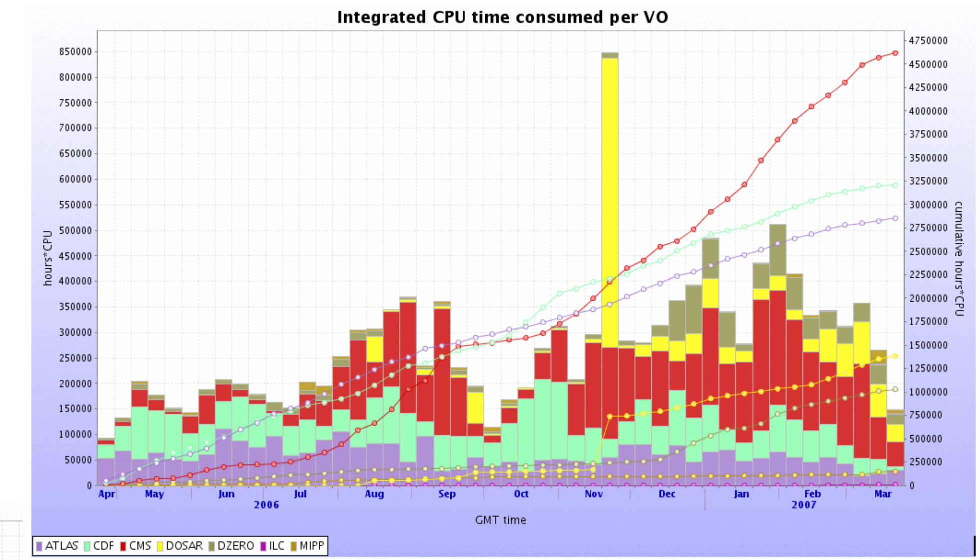
Assumes flat funding, particular allocation of funds and changes in technology.

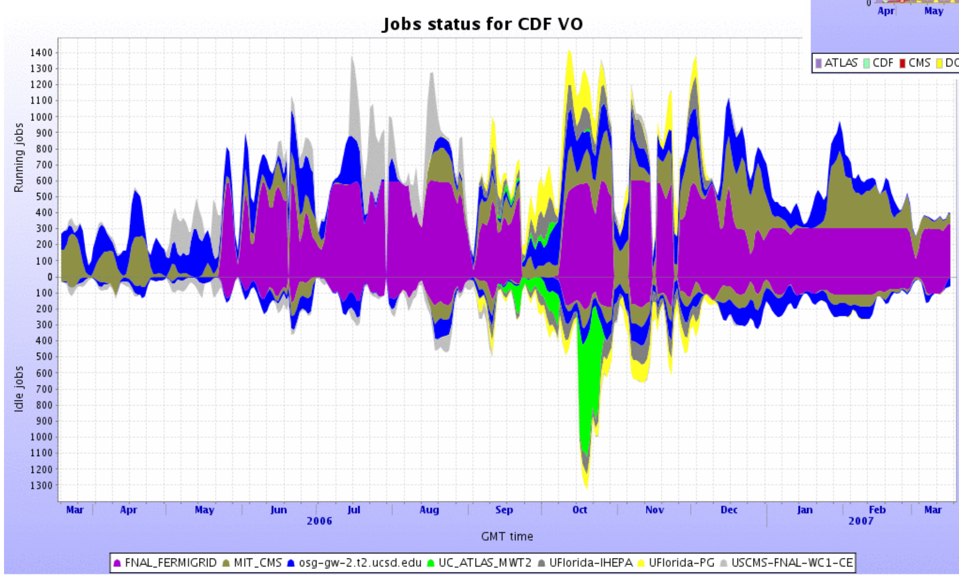
		Actual		Requirements		
	Fiscal Year	2006	2007	2007	2008	2009
CPU (THz)	Fermilab	4.8	7.6	7.6	7.8	7.9
	On-site contributions	1.7	1.7	1.7	1.7	1.7
	Remote (dedicated)	2.3	2.3	2.3	2.3	2.3
	Opportunistic		1.0	1.0	5.0	13
	Total	8.8	13	13	17	25
Disk (PB)	Fermilab	0.6	0.6	0.6	0.9	1.6
	On-site contributions	0.2*	0.1	0.1	0.1	0.1
	Remote	0.1?	0.1?			
	Total	0.9?	0.9?	0.7	1.0	1.7
Volume on tape, Oct. 1, 2006 (PB)		1.5	1.6	3.8	6.4	10.3

* 50% retires at end of FY2006

CDF utilization of OSG

CDF not the largest user, but is a significant consumer of OSG resources





Significant opportunistic use of OSG sites, although majority of CPU used is in Fermigrid

- Can we obtain needed resources opportunistically?
  - Already OK this year
  - Anticipated demand is 2% (3%) of LHC total in FY2008 (FY2009)

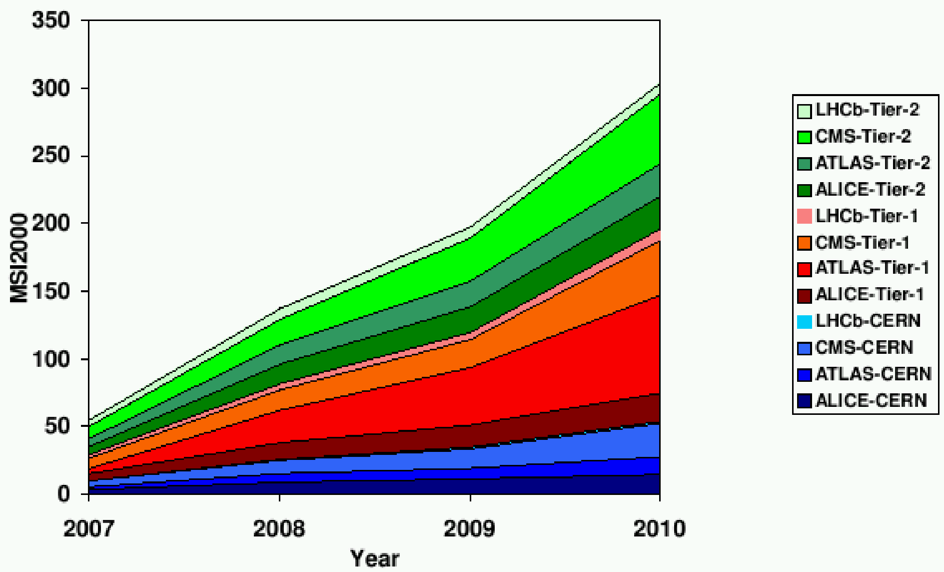


Figure 1 CPU requirements at the Tier-0, Tier-1 and Tier-2 centres time period is from 2007 to 2010.

Computing demand reduction strategies

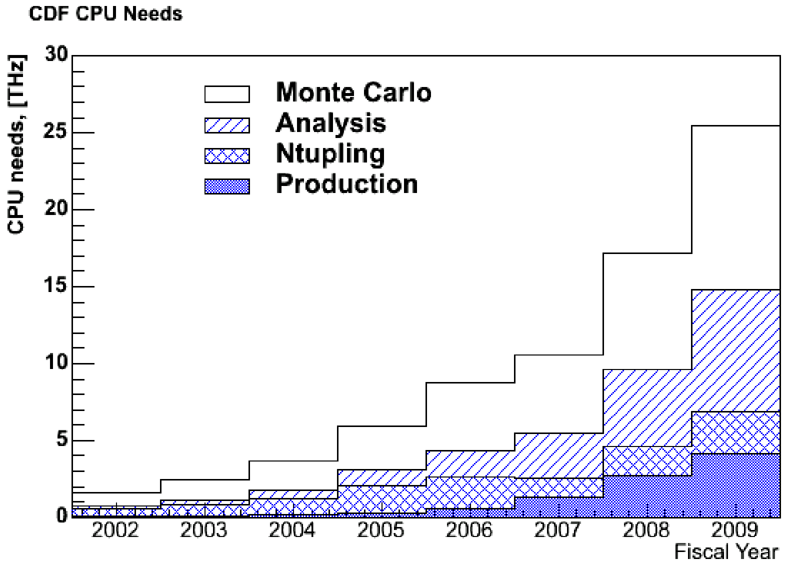
- Projections assumed particular allocation of funding
  - Can alter allocations, adjust trade-offs, tune parameters of the computing model
- Continue to improve utilization efficiency
  - Doing this already (e.g., V7 offline release)
- Prioritize computing tasks
  - Defer processing some data
  - Reduce MC statistics

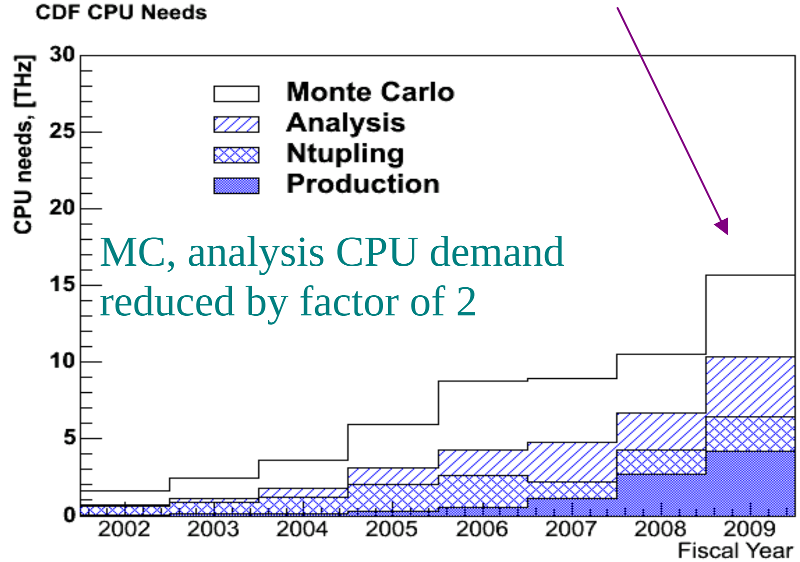
Could imagine reduction of a factor of two in combined CPU for MC, ntupling and analysis, but at some cost to the physics

Computing demand reduction strategies

- Prioritize computing tasks (cont'd)

Opportunistic demand in this scenario= 1.7% of LHC computing





- Logging rate may not double by FY2009 (not a strategy…)
  - Depends upon improvements in trigger algorithms (DAQ already capable of 60 MB/s)

Many handles on the problem

Summary

- Offline systems are working well
  - Provide rapid turn-around between data-taking and physics results
  - Current efforts aimed at improving efficiency, reducing operational load
- Judiciously expanding use of grid, opportunistic computing
  - Successfully meeting requirements so far
  - Have flexibility and options available for the future
- Expect effort required to operate offline systems to fall
  - Also expect effort contributed by experiment to fall
  - More systems likely to become responsibility of CD
    - Will have sustained effort from CD
- Confident of continued success through end of data taking

The end

Backup slides

Details of CDF security incident

1) CIAC# 518276 03/29/2006 Impact: CDF

Alert: Mail message from CIAC warning that a Fermilab node was participating in an IRC chat room that is regularly monitored by US-CERT.

Status: Machine was found to be running a web server with a version of PHP which had a known vulnerability. Intruder was able to install and run an IRC client. No local access to the machine was gained. Node had a legacy exemption to allow running web server.

Root Cause: Vulnerable version of utility software installed. Misconfiguration of IPTABLES blocked local security scan which would have identified vulnerability.

No impact on CDF Online major application (back)

Collaboration Resources

	CY 2007	2008	2009
US FTE	222	162	127
Non US FTE	170	135	109
Total US + Non US	392	297	236
Post Doc's	101	73	53
Students	147	102	77

- Collaboration members available in units of FTE
  - Delay in LHC turn-on and the success of the Tevatron has resulted in many more FTE's available for CDF
    - 30% increase over 2005 estimates
  - Students and post doc's decline at similar rate
  - Total head count remains steady at ~610 physicists

Collaboration Resources Needed

	CY 07	CY 09
Detector Ops	50	45
Offline *	26	20
Algorithms	32	21
Management	10	10
Total	118	96
Resources Available	392	236
FTE for Physics	392 – 118 = 284	140

- More than enough people to run the experiment and do physics!